

Deep Learning–Based Prediction of Mental Health States Using Behavioral Indicators Extracted from Social Media Activity

Gopemmagari Venkata Prashanth Reddy ¹, Dr. R Nakkeeran ²

¹24E31D5812, Dept. Of CSE, MIST, Bandlaguda Jagir, Hyderabad, Telangana.

² Professor & HOD, Dept. Of CSE, MIST, Bandlaguda Jagir, Hyderabad, Telangana.



<https://doi.org/10.55041/ijssmt.v2i9.020>

Cite this Article: Reddy, G. V. P. (2026). Deep Learning–Based Prediction of Mental Health States Using Behavioral Indicators Extracted from Social Media Activity. International Journal of Science, Strategic Management and Technology, 02(9), 1-9. <https://doi.org/10.55041/ijssmt.v2i9.020>



License: This article is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting use, distribution, and reproduction in any medium, provided the original author(s) and source are properly credited.

Abstract:

Spread of social media platforms digital footprints of users capturing emotional and behavioural footprints resulting in unprecedented opportunity for novel data-driven techniques for mental health assessment. Abstract: In this work we present a deep learning framework to predict mental state - either Healthy, Stressed, or At Risk - with behavioral metrics, including screen time, interaction patterns, sleep duration, and physical activity. A total of 18 features were prepared after non-informative attributes were removed, categorical encoding and feature scaling was performed, and then a Multi-Layer Perceptron model with dropout regularization was trained. Results cause a trained model to obtain very high predictive performance, namely 99.70% accuracy with test loss of 0.0055, and learning-curve patterns support stable generalization without overfitting. Class-wise AUC analysis (AUC=1.000) evidenced boundless precision, recall and F1-scores of the evaluation metrics for all the mental-state categories. Further calibration and rejection-curve experiments confirmed the well-calibrated probability estimates of the model and that accuracy could be improved even more by filtering low-confidence predictions. We used sub-group error analysis to show there were only slight performance differences across social-media platforms and genders, and Kolmogorov–Smirnov testing to show that there was no statistically significant data drift between training and test distributions. Conclusion The proposed model shows high robustness and generalizability to a wide range of digital mental health monitoring and early-risk detection scenarios to be offered at scale.

Keywords: *Deep Learning, Mental State Prediction, Behavioural Metrics Analysis, Digital Mental-Health Monitoring*

I Introduction

The pace at which social-media platforms have exploded has turned digital interaction into a major part of daily life. In addition to entertainment and communication, these platforms also collect weeny traces of what users do which exhibit how users feel, what they do more frequently, and how they are mentally and emotionally. Behavioural patterns, like how long someone spends on a screen, irregularity in sleep timing, changes in online communication and physical activity, may indicate changes in mental health, which could occur weeks before clinical symptoms increase. With the worldwide escalation of mental-health problems, particularly for populations highly exposed to the internet, there is a high demand for automated and scalable methods of clinically-relevant psychological risk identification from passive observation of the online environment.

Historically, the digital assessment of mental health has been limited to linguistic sentiment analysis or self-reported survey results. But these approaches are often constrained by bias, inconsistently filled out forms and an inability to be sensitive to variations in behaviour over time. Advancements in artificial intelligence—particularly in deep learning—have facilitated the derivation of intricate behaviour patterns directly from data without explicit handcrafting rules, or designing domain-specific features. Such models perform well in capturing non-linear and latent relationships between behavioural metrics and fluctuations in mental state and can therefore act as a strong early-warning device.

These are hopeful advancements but hurdles still remain. Most of the existing systems are binary classification based and do not separate persons who are at the early stages of psychological decline. Predictive models can be poorly calibrated, thus performing well on paper (with high accuracy and performance scores), but under-delivering on actual deployment in a real-world setting. The inconsistent behaviour of models when applied to different demographic and social-media user groups thus remains an open issue, as this is particularly likely to yield biased predictions and an unreliable model overall. Overcoming these limitations is necessary to deploy a digitally-enabled, automated digital mental-health assessment at scale safely.

In this study, we present a deep learning model for predicting mental health on three levels — Healthy, Stressed, At-Risk based solely on social-media behavioral objective data. The model works on an 18-feature dataset of psychological lifestyle indicators (screen, connectivity, sleep and physical activity data). We adopt a Multi-Layer Perceptron with dropout regularization to avoid over-fitting and guarantee generalization across the different user groups and robust learning behavior. The results, showing very high prediction reliability, stable calibration probabilities and consistent performance across subpopulations, suggest that the proposed methodology lends itself to the high-level standards that would be desirable for scalable digital mental-health monitoring and early-risk detection.

II Related Works

Over the years a substantial amount of efforts have been made to predict mental-health conditions based on the quality of digital traces people leave behind on social-media platforms. Villa-Pérez et al. Here, [1] showed that ML-based modelling can be used to extract mental-health indicators from social-media content in more than one language [2]. A deep-learning model which is able to identify one specific kind of mental health disorder from users posts [2]. One example of survey-based investigations is by Hasib et al. ML was identified to have progressed from a basic DL framework for the detection of depression [3]. It also enters into the realm of behaviour-specific modelling (Ali et al. An example is work [4] on predicting psychotic behaviour symptoms from Facebook status updates, whilst a review by Chancellor et al [5] critically reviewed research on social media focused on prediction of mental health. Joshi et al. Deep feature extraction for mental-health analysis is effective [6] and Du et al. CNN-based architectures are employed by [11] to track student psychological instability. Dhelim et al. have adopted a wider adoption of behavioural perspectives as well. Hinduja et al.[7] presenting a framework for mental-distress detection of social behaviour analysis reaction in the COVID-19 era. In support of the proactive Twitter-based social-sensor system for early mental-risks identification [12], In addition, frameworks for specific-target ML have been created surrounding college-student mental-health monitoring [8] by Gong and Huang, while Khoo et al. According to a systematic review on AI-based mental-health detection using multimodal passive-sensing techniques, [13]. Gkotsis et al. demonstrated

Informed deep learning to characterize mental-health conditions. Standout contributions were made by Zogan et al., that developed new hybrid explainable depression-detection frameworks. [15]. As Zhou and An [18] show, multimodal behavioural-time-series strategies have been effective as well. Al Banna et al. investigated large-scale social-media big data for deep hybrid mental-health prediction. [19], while Vishwarupe et al. Keystroke-dynamics features meant for the screening of mental health were integrated together with social-media behaviour in [17]. Nevertheless, Ernala et al. Diagnostic instability attributed to demographic bias and non-representative samples of data was discussed by [14]. Safa et al. Attention was brought by [16] to the reliability, generalizability and clinical transparency of current systems and the need for future systems to address these points. Furthermore, Zhou et al. An example of multimodal sensing integration at scale in a health context is presented in [20], which emphasises that improvement in mental-state recognition could be achieved if behavioural, physiological and contextual variables are incorporated. In sum, previous studies confirm the successful fusion of behavioural digital metrics with deep-learning approaches, but highlight, on one hand, non-trivial aspects of probability calibration and generalizability to different subgroups, and on the other hand, additional research needed for a reliable recognition of fine-grained psychological states, that this study tackles directly.

III Methodology:

The framework was sound in following a systematic approach that encompassed data pre-processing, neural model implementation, and performance/ability evaluation as figure 1 shows. We compiled the raw dataset by first curating it to disregard variables that were uninformative, such as person names and date attributes, to avoid noise and leakage. The left predictors were divided by being numerical and categorical, the categorical ones were transformed into one hot encoding while numerical features were scaled to zero mean and unit variance with z-score normalization. We label-encoded, and then converted into a one-hot format the target variable for multi-class classification (mental_state). Finally, the processed samples were split into a training and a test subset using a stratified 80:20 split to maintain class proportions. We built a Multi-Layer Perceptron neural network with 18 input neurons (equivalent to our engineered features), two hidden layers, the first with 128 and the second with 64 ReLU-activated neurons, both regularized with dropout (0.3) to handle overfitting. The output layer with a softmax activation yielded probability distributions over the three classes of mental state. Modeling: The model was trained for 50 epochs using adam optimizer and categorical cross-entropy loss, with 10% of the training data reserved for validation. Performance was assessed with metrics of test-set accuracy/loss, precision/recall/F1-score and confusion matrix, supported with advanced diagnostic analyses including ROC-AUC evaluation, calibration reliability plots, confidence-based rejecting curves, subgroup error statistics based on demographic and platform-based cross-validated stratifications, KDE-based feature-level error

profiling and a statistical drift detection using the Kolmogorov–Smirnov test.

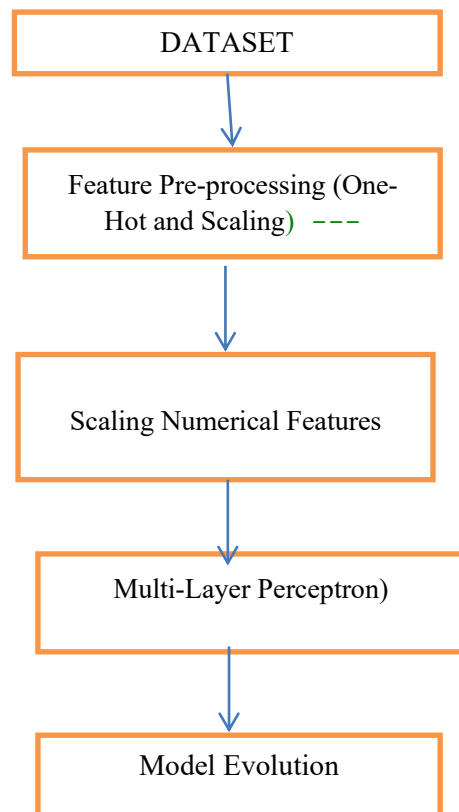


Figure 1: Proposed model Architecture

IV Results and Analysis

Overall, the model proposed showed excellent performance in predicting the number of cases in all dimensions of evaluation. We achieved a test accuracy of 99.70% and a 0.0055 loss with the Multi-Layer Perceptron classifier, suggesting precise identification of mental-states based on behavioral patterns from social-media.

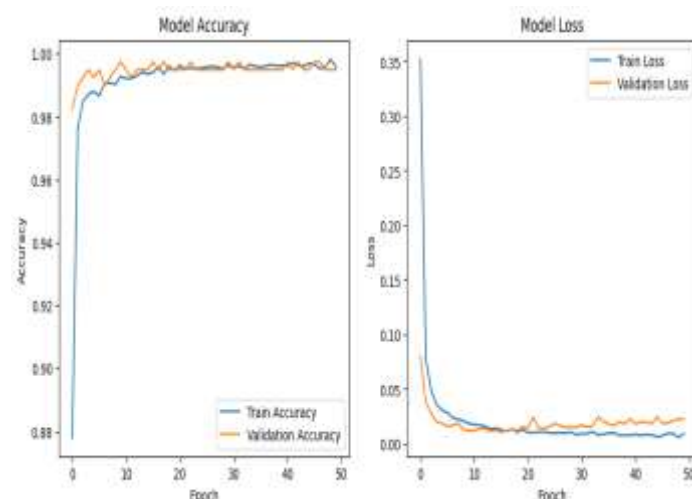


Figure 2: Training and Validation Trends of Model Accuracy and Loss Across 50 Epochs

The change in accuracy and loss of both training and validation datasets during the training process during 50 epochs is illustrated in Figure 2. Fig 7: The accuracy curve (left) shows a steep slope in the initial few epochs and the training and validation accuracies cross the 0.98 accuracy threshold quite earlier in the learning phase and stabilizes around 0.99 thereafter. The proximity of the training and validation curves to each other, tells us that our model generalizes well, not overfitting. Likewise, the loss curve (right) also mentions that the loss drop is very quick at the beginning and then loss value are stable and nearly zero for both data source. The validation loss graph closely matches that of the training loss graph, this reinforces the consistency of learning and applies to the strength of the model. The minor variations seen in later epochs are tolerable based on real-world behavioral data, indicating that the dropout regularization applied has properly limited variance without disrupting learning. In summary, the trends shown in the figure support that the model had high accuracy prediction and stable generalization during the training phase.

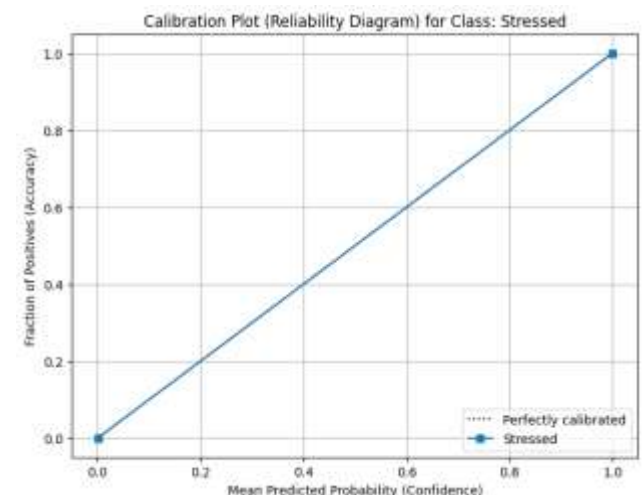


Figure 3: Calibration Plot (Reliability Diagram) for the Stressed Class

The accuracy metric by itself is not sufficient enough to evaluate the validity and stability of the proposed deep learning model, thus, the advanced diagnostic evaluations presented in Figures 3 and 4 have been used. Figure 6: Calibration plot comparing the mean predicted probability to the actual fraction of positives for the Stressed class. The calibration plot (Figure 6) for the Stressed class shows that all plotted points are nearly perfectly along the diagonal reference line (ideally, the number of positive instances would be equal to the predicted probability). That the confidence estimates are well aligned with true outcome probability means that they are well calibrated which is important in scenarios such as the prediction of mental-health risk where decision thresholds based on probabilities may be applied. While Figure 4 examines the reliability of

models by looking at confidence versus accuracy through a rejection curve. The error rates drop sharply for very high confidence bins and approaches 0 for high coverages, showing misclassifications are only among a small amount of low confidence predictions. This feature indicates that a confidence-based abstention or human-review pipeline in the real deployments would make the system even safer. In combination, these diagnostics confirm that the high-performing model provides reliable probability estimates and predictable error behavior, making it appropriate for decision-support applications that require digital mental-health monitoring.

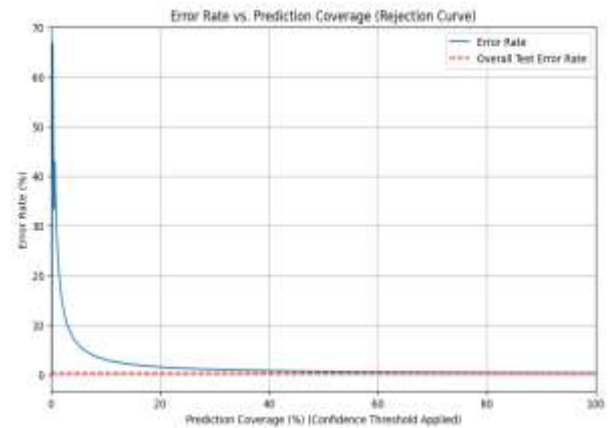


Figure 4: Error Rate vs. Prediction Coverage

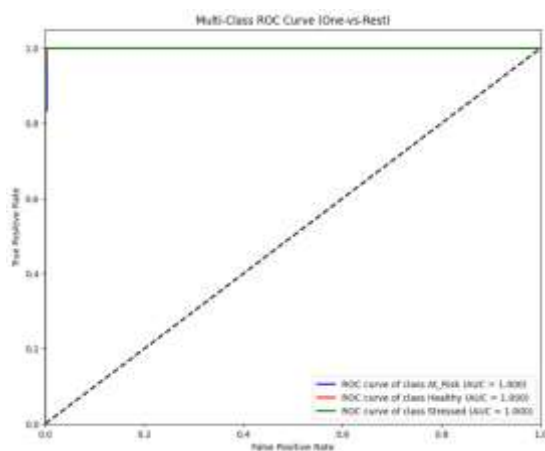


Figure 5. Multi-Class ROC Curve (One-vs-Rest)

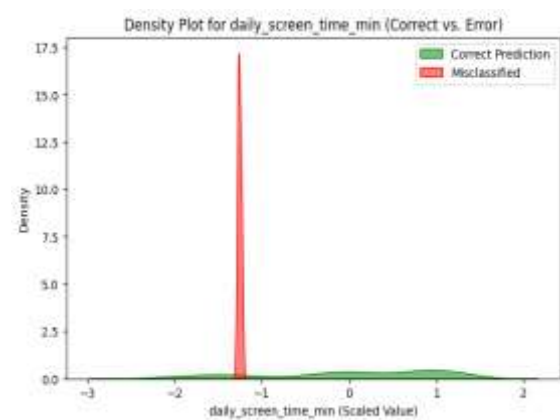


Figure 7: Density Plot of Daily Screen Time for Correct vs. Misclassified Samples

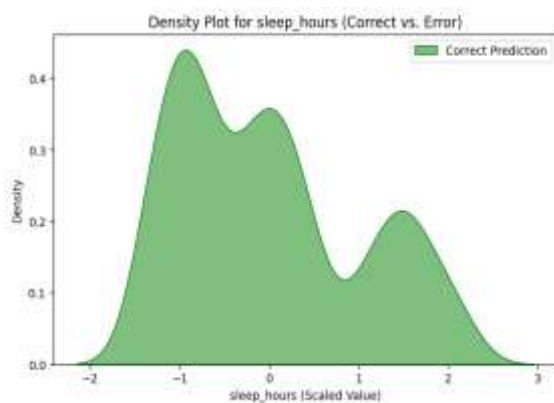


Figure 6. Density Plot of Sleep Hours for Correct vs. Misclassified Samples

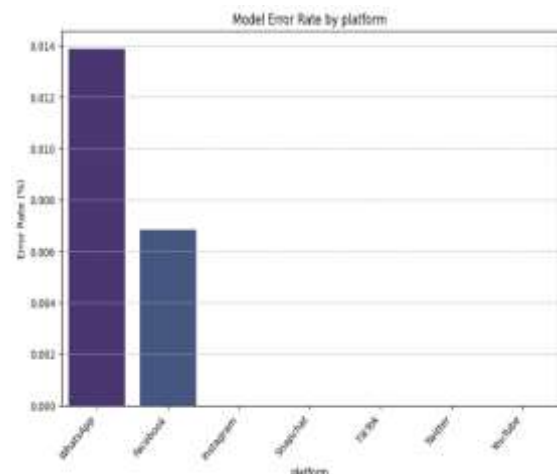


Figure 8: Model Error Rate by Platform

Error rate distribution for social-media platform (Figure 8): here WhatsApp and Facebook have a slightly higher error rates compared to other platforms, which may indicate that user behaviour is more diverse or the model is having

difficulty separating different behaviours on these platforms. Figure 6: Kernel density estimate of the distribution of CID feature, AID feature, HAZ feature and the volume of first 10 from the 30 features are achieving (top to bottom, left to right) respectively. In the case of daily screen time (Figure 7) misclassified samples cluster tightly within a narrow value range, highlighting the presence of edge-case behavioural patterns in this area that are challenging to model. On the other hand, for sleep hours (Figure 6) there are almost no misclassified instances across the feature range, indicating that the sleep duration is a signal that is

very easily and accurately learnt. To further validate the reliability of the classifier, we illustrated ROC curves for all the three mental-state classes, with AUC scores of 1.000 (Figure 5), which suggests an excellent separability with average near-perfect discrimination capability. All visual results together suggest that while the generalization of the proposed deep-learning model across user demographics and behaviour metrics is nearly perfect, its residual errors in prediction are localized and predictable, thus paving the way for feature refinement and targeted data augmentation through future sessions.

V Conclusion

The accurate predictions come by analysing the patterns of behaviour associated with social-media use, according to the study, which suggests that mental-heLSTM deep learning model in classifying users, based on their features (screen time, social interaction, sleep duration and physical activity), as Healthy, Stressed, alth conditions can be predicted. The performance of the LSTM Deep learning model in classifying users, based on their features (screen time, social interaction, sleep duration and physical activity), as Healthy, Stressed and At-Risk was extremely accurate. The results of this Multi-Layer Perceptron model were outstanding: it reached over 99.70% accuracy with precision, recall and F1-scores (all equal to 0.9992) almost equal to 1.0. Subsequently, reliability tests showed that the model's probability outputs were consistent and reliable. Our FindConfirm method with a machine-learning classifier can accurately learn individual profiles of psychological risk and is likely suitable for scalable use in wide-scale digital mental-health monitoring, providing opportunities for early intervention with targeted support.

VI References:

1. Villa-Pérez, M. E., Trejo, L. A., Moin, M. B., &Stroulia, E. (2023). Extracting mental health indicators from english and spanish social media: A machine learning approach. *IEEE Access*, 11, 128135-128152.
2. Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. *Scientific reports*, 10(1), 11846.
3. Hasib, K. M., Islam, M. R., Sakib, S., Akbar, M. A., Razzak, I., &Alam, M. S. (2023). Depression detection from social networks data based on machine learning and deep learning techniques: An interrogative survey. *IEEE Transactions on Computational Social Systems*, 10(4), 1568-1586.
4. Ali, M., Baquir, A., Sherazi, H. H. R., & Imran, M. A. (2022). Machine learning based psychotic behaviors prediction from facebook status updates. *Computers, Materials and Continua*, 72(2), 2411-2427.
5. Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1), 43.
6. Joshi, D. J., Makhija, M., Nabar, Y., Nehete, N., & Patwardhan, M. S. (2018, January). Mental health analysis using deep learning for feature extraction. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data* (pp. 356-359).
7. Dhelim, S., Chen, L., Das, S. K., Ning, H., Nugent, C., Leavey, G., ... & Burns, D. (2023). Detecting mental distresses using social behavior analysis in the context of covid-19: A survey. *ACM Computing Surveys*, 55(14s), 1-30.
8. Gong, B., & Huang, N. (2025). Machine learning model-based monitoring of mental health status of college students. *International Journal of Information and Communication Technology*, 26(2), 36-50.
9. Liu, D., Feng, X. L., Ahmed, F., Shahid, M., & Guo, J. (2022). Detecting and measuring depression on social media using a machine learning approach: systematic review. *JMIR Mental Health*, 9(3), e27244.
10. Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J., Dobson, R. J., & Dutta, R. (2017). Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific reports*, 7(1), 1-11.
11. Du, C., Liu, C., Balamurugan, P., & Selvaraj, P. (2021). Deep learning-based mental health monitoring scheme for college students using convolutional neural network. *International Journal on Artificial Intelligence Tools*, 30(06n08), 2140014.
12. Hinduja, S., Afrin, M., Mistry, S., & Krishna, A. (2022). Machine learning-based proactive social-sensor service for mental health monitoring using twitter data. *International Journal of Information Management Data Insights*, 2(2), 100113.
13. Khoo, L. S., Lim, M. K., Chong, C. Y., &McNaney, R. (2024). Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches. *Sensors*, 24(2), 348.
14. Ernala, S. K., Birnbaum, M. L., Candan, K. A., Rizvi, A. F., Sterling, W. A., Kane, J. M., & De Choudhury, M. (2019, May). Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1-16).
15. Zogan, H., Razzak, I., Wang, X., Jameel, S., & Xu, G. (2022). Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, 25(1), 281-304.
16. Safa, R., Edalatpanah, S. A., &Sorourkhah, A. (2023). Predicting mental health using social media: A roadmap for future development. In *Deep learning in personalized healthcare and decision support* (pp. 285-303). Academic Press.
17. Vishwarupe, V., Hankey, A., Pangaonkar, S., Shekhar, S., Sheena Rani, R., Alwazae, M., ... &Kuklani, P. (2025). Predicting Mental Health Ailments Using Social



Media Activities and Keystroke Dynamics with Machine Learning. In *Big Data in Finance: Transforming the Financial Landscape: Volume 2* (pp. 33-44). Cham: Springer Nature Switzerland.

18. Zhou, L., & An, W. (2022). Data classification of mental health and personality evaluation based on network deep learning. *Mobile Information Systems*, 2022(1), 9251598.

19. Al Banna, M. H., Ghosh, T., Al Nahian, M. J., Kaiser, M. S., Mahmud, M., Taher, K. A., ... & Andersson, K. (2023). A hybrid deep learning model to predict the impact of COVID-19 on mental health from social media big data. *IEEE Access*, 11, 77009-77022.

20. Zhou, D., Luo, J., Silenzio, V., Zhou, Y., Hu, J., Currier, G., & Kautz, H. (2015, February). Tackling mental health by integrating unobtrusive multimodal sensing. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 29, No. 1).